The stochastic model of the impact of context factors to educational results of Tomsk school graduates

Yuliy Ya. Katsman, Alexey V. Lepustin
Computer Engineering Department
Tomsk Polytechnic University
Tomsk, Russia
kim@tpu.ru

Boris V. Ilyukhin, Elena V. Lepustina
Radiosystems Chair
TUSUR
Tomsk, Russia
bvi@ege.tomsk.ru

Zhanna N. Zenkova
Applied Mathematics & Cybernetics
Tomsk State University
Tomsk, Russia
thankoff@fpmk.tsu.ru

***Abstract*—The article describes the factor model of Tomsk region schools functioning which was developed and investigated using the STATISTICA system. The constructed model describes an impact of different variables (context factors) on educational results of Tomsk school graduates. At the preliminary stage the most significant variables were determined, the exploratory data analysis was made using the method of principal components. There were formulated 3- and 4-factor models using Kaiser-Guttman's and Cattell's Criteria. Factor rotation with Varimax method allowed to interpret factor loadings clearly. In this work there were investigated the quality of the factor model for different types of schools such as urban, country, ungraded ones.**

Keywords—Factor model, factor loading, determination coefficient, dispersion, correlation, matrix, criterion, method of principal components, eigenvalues.

I. INTRODUCTION

Nowadays in Russian Federation the most standardized education quality assessment procedure is the Unified State Exam (USE). The education system in Russia is not a closed system, so the individual achievements of school students are influenced by the so-called context factors: the socio-economic level of the territory, educational qualification of the parents, etc. [1, 2]. The investigation [3] discovered the relation between students' results as well as school administrators' and teachers' attitude to the USE, teachers' qualification and age, complex social factors of school functioning, etc. This article takes aim to reveal different factors, which are most significant for educational results. The other objective is to improve the methods of the different types of schools comparison [3]. The method, proposed at [2], became the basis to investigate Tomsk school USE-results and to develop stochastic model.

The input data is the matrix with ~180 variables (factors) and more than 200 lines (schools of different types: urban, country and ungraded). For each line (school) about 180 context factors were formed, such as "Total amount of teachers", "Number of families with both parents unemployed", "Total amount of students, having "Good" and "Excellent" marks at the secondary school", etc. The matrix of conjugate correlations with dependent variables (results of the USE) was used to develop the most significant context factors. In previous papers the multivariate linear regression models with different sets of independent variables were constructed [4, 5]. The optimal dimension (from 12 to 1) was selected to provide the model quality (all the regression equation coefficients are significantly deviated from 0, the adjusted coefficient of multiple determination (R^2_{adj}) is maximal). This study showed that for most schools and different sets of context factors the 3-dimension model is optimal, though maximal (R^2_{adj}) were less than 0.5 that proves model to be insufficient. Further investigation was made in the framework of factor analysis.

The main aim of the factor analysis is to reduce the model, e.g. to decrease the number of considering variables. This reduction is attained by the extracting hidden general factors, which cannot be measured directly. These factors explain the relations between observed characteristics (variables), so, instead of original set of variables, we can analyze the extracted factors (the quantity of extracted factors is considerably less than original set of interconnected variables). The advantage of the factor models application is the fact that these models are working well not only in technical systems, but also in medical, biological, social and other systems [6].

In this paper the factor analysis model is described in more detailed way. The original set of data is presented as a matrix $X = [x_{ji}]$, $i = \overline{1,n}$, $j = \overline{1,N}$, where N is a number of considered subjects, n is a quantity of measured variables. Suppose, that in the factor model each element x_{ji} should be the result of the influence of few (m) hypothetic general factors (m<<n) and a character factor [7]:

$$x_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \ldots + a_{jr} f_{ri} + \ldots + a_{jm} f_{mi} + d_j v_{ji}, \qquad (1)$$

where α_{jr} is a weight coefficient (or load) for j^{th} variable of r^{th} general factor; f_{ri} is meaning of r^{th} general factor of i^{th} subject of inquiry; d_j is a weight coefficient (or load) for j^{th} variable of j^{th} character factor; v_{ji} means the j^{th} character factor of i^{th} subject of inquiry; $j = \overline{1,n}$, $i = \overline{1,N}$, $r = \overline{1,m}$, $m << n$.

Taking into account that original data set $X = [x_{ji}]$ contains the variables of different dimensions, we have to

standardize the matrix elements and than investigate the factor model:

$$y_{ji} = (x_{ji} - \overline{X}_j) / S_j \qquad (2)$$

where $\overline{X}_j$ – is a sample mean of j^{th} variable (character); S_j is a sample standard deviation for j^{th} variable.

Finally, we considers the model

$$y_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \ldots + a_{jr} f_{ri} + \ldots + a_{jm} f_{mi} + d_j v_{ji}, \qquad (3)$$

where a_{jr} is an unknown coefficients (factor loads); $d_j v_j$ is the remainder (residual), or residual specific factor. Now, the problem is to estimate optimally the factor loads a_{jr}.

In this research the method of principal components is used, the minimum of the divergence between original character covariance matrix and the matrix, derived after factor loads estimation was used as a criterion of optimality. In this case, the measure of the variance of these matrixes is the Euclidean norm of their divergence [8].

All stages of the model development, research, analysis, presentation of the obtained results were implemented using the STATISTICA software package.

At first stage of this investigation (as for regression analysis) the correlation matrix to discover variables with the correlation coefficient considerably deviated from 0 is calculated. The data analysis included all schools of the Tomsk region. The results of this analysis allowed to distinguish 16 most important variables (Table I).

TABLE I. THE VARIABLES, USED IN THE MODEL

№ of variable	Name of variable
1	The share or part of schools with psychologists + speech and language pathologists
2	The part of schools with pedagogues of additional education
3	The part of declining families
4	The part of families with both parents working
5	The part of families with both parents having higher education
6	The part of families with one of the parents having higher education
7	The part of families living in socially dangerous conditions
8	The part of students having a police records (or other services)
9	Total amount of students studying profile programs at the 10-11th grades in 2011-2012 years.
10	Total amount of classes having profile programs at the 10-11th grades in 2011-2012
11	Total amount of students having "Good" and "Excellent" marks at basic school in 2011-2012
12	Total amount of students, having " Good" and "Excellent " marks at the secondary school in 2011-2012
13	The percent of basic part fulfillment of the USE (Russian language)
14	The average score in the USE Russian language
15	The percent of basic part fulfillment of the USE (Mathematics)
16	The average score in the USE Mathematics

The table I contains both types of context factors (variables) such as initial factors ("The average score in Mathematics at the USE", "Number of participants and winners of the contests", etc) and transformed factors ("The part of schools with psychologists + logopedists + defectologists", "The part of students having a police records (or other services)", etc). Actually, the variable "The part of psychologists + logopedists + defectologists" may be equal in different schools, in contrast to the variable "Total amount of teachers". Obviously, it is correct to use the relative (normalized) values of the context factors instead of the absolute values (i.e., to normalize the initial factors by the total amount of students or the average amount of teachers in school, etc.).

Formally, these variables characterize three groups of indicators:

1. Qualification and innovation indicators of the school;

2. Social and economical conditions;

3. Educational results of students (including results of the USE in mathematics and the Russian language).

There was pointed [4] that the influence of the context factors depends on the school type that is urban, country, ungraded school. According to this, the characteristics of the factor models, developed for different types of schools are considered.

II. RESULTS AND DISCUSSION

A. *The exploration analysis*

It is supposed at the stage of the exploratory data analysis that the number of factors to extract m is equal to the quantity of variables n. The method of principal components [9] supposes that the variance of each variable equals 1. This hypothesis is reasonable if we remember that the factor model (3) is developed based on standardized original data (2). So the total variance equals the sum of variable variances (16 in our case). Each factor has the corresponding variance. The variances of distinguished factors are called the eigenvalues. Using the method of principal components [9] we get the eigenvalues of factors (Table II).

TABLE II. THE EXPLORATORY DATA ANALYSIS

Value	Eigenvalues (Factor) Extraction: Principal components			
	Eigenvalue	*% Total variance*	*Cumulative Eigenvalue*	*Cumulative %*
1	6,057701	37,86063	6,05770	37,8606
2	1,918289	11,98930	7,97599	49,8499
3	1,271599	7,94749	9,24759	57,7974
4	1,154225	7,21391	10,40181	65,0113
5	1,019095	6,36935	11,42091	71,3807
6	0,750309	4,68943	12,17122	76,0701
7	0,709055	4,43159	12,88027	80,5017
8	0,640438	4,00274	13,52071	84,5044
9	0,568002	3,55001	14,08871	88,0545

Value	Eigenvalues (Factor) Extraction: Principal components			
	Eigenvalue	*% Total variance*	*Cumulative Eigenvalue*	*Cumulative %*
10	0,512400	3,20250	14,60111	91,2570
11	0,467288	2,92055	15,06840	94,1775
12	0,380023	2,37514	15,44842	96,5526
13	0,253998	1,58749	15,70242	98,1401
14	0,216341	1,35213	15,91876	99,4923
15	0,062565	0,39103	15,98133	99,8833
16	0,018671	0,11670	16,00000	100,0000

[a] Description of the table II: Column 1 – factor number, 2 – factor variance (eigenvalue); 3 – percent of total variance; 4 – cumulated variance; 5 – percent of cumulated variance.

According to calculations we can conclude that the first factor gives the ~38% of the total variance, the second does 12% and so on. The bigger the factor number is, the smaller its contribution to total variance becomes (i.e. 16th factor gives less than 2% of total variance). Also we can see that the second factor gives 3 times less contribution than the first one, but the 3-5th factors' contributions are approximately equal. So, when we know the dispersions corresponding to each factor, we have to decide how many factors are optimal for the multifactor model.

To solve this problem the Kaiser-Guttman's Criterion [10] is used. This criterion implies taking into account only those factors that have the eigenvalues more than 1. So we can ignore factors with corresponding variance that is less than variance of one variable. In this case we have to limit our model to 5 factors, giving 71% of total variance. This multifactor model is quite complicated (5 factors) and not adequate (describes just 2/3 of total variance).

But it would be better if we could develop simplified model (1-3 factors), described not less than 80-85 % of total variance. To fix the optimal number of factors we use Scree plot or Cattell's criterion [11]. The idea of the criterion is to plot the dependence between eigenvalue and factor number. It is reasonable to limit the number of factors where the decreasing the eigenvalues from left to right will maximally slow down. See the plot (Fig. 1)

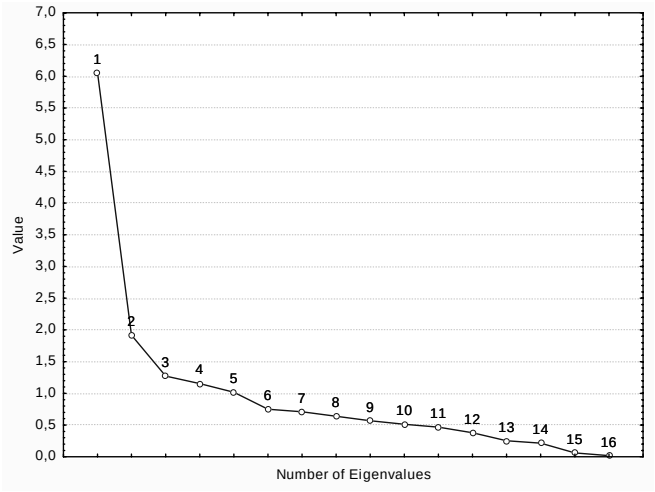

Fig. 1. Plot of Eigenvalues

Analyzing this plot we can see that we have to limit the number of factors to 5-6, as at these points the shedding is slowing down considerably.

To continue the exploratory data analysis we estimate total factor loadings, defined the correlation between variables and corresponding factors. Partial results of factor loads without the factor rotation are presented at the Table III. Also there are marked correlations bigger than 0.7 we use their numbers and the numbers of school character group for this variable instead of long variables names (column 1). Last two rows contain factor variance and its part at the total variance correspondingly.

We can see that the first and the second factors have bigger correlation coefficients than other factors, besides, the bigger the factor number is, the smaller is the correlation coefficient, but decreasing is not very rapid.

TABLE III. TOTAL FACTOR LOADINGSS OF EXPLORATORY DATA ANALYSIS

Var #	Factor Loadings (Unrotated Factors) Extraction: Principal components (Marked loadings are >,700000)					
	Factor_1	*Factor_2*	…	*Factor_5*	…	*Factor_16*
1 {1}	-0,2815	0,5083		-0,1609		0,0009
2 {1}	-0,2489	0,5022		-0,0316		-0,0008
3 {2}	-0,4831	0,2331		-0,0666		0,0026
4 {2}	**-0,7163**	0,3325		-0,0621		-0,0055
5 {2}	**-0,7388**	0,2739		-0,1103		0,0063
6 {2}	-0,5610	0,3352		-0,0732		0,0010
7 {2}	0,3655	-0,0189		**-0,7393**		-0,0008
8 {2}	0,4680	-0,0965		-0,6246		-0,0005
9 {1}	**-0,7396**	0,0771		-0,0902		0,0081
10 {1}	-0,4659	-0,0795		-0,0294		-0,0021
11 {1}	**-0,7300**	0,3400		0,0053		-0,0099
12 {1}	-0,6324	0,0975		-0,0595		0,0047
13 {3}	**-0,7855**	-0,3697		-0,0677		-0,0127
14 {3}	**-0,7859**	-0,3813		-0,0579		0,0167
15 {3}	**-0,7029**	-0,5631		-0,0653		0,0916
16 {3}	**-0,7287**	-0,5559		-0,0704		-0,0978
Expl.Var	6,0577	1,9183		1,0191		0,0187
Prp.Totl	0,3786	0,1199		0,0637		0,001167

B. Three-factor model development

Using the exploratory data analysis results and Cattell's and Kaiser-Guttman's Criterions, we try to develop simple (three-factor) model and estimate its quality. The factor loadings for this model are presented in the Table IV.

TABLE IV. FACTOR LOADINGSS FOR THREE-FACTOR MODEL

Variable	Factor Loadings (Unrotated) (Factor) Extraction: Principal components (Marked loadings are >,700000)		
	Factor_1	*Factor_2*	*Factor_3*
1 {1}	-0,2815	0,5083	0,1658
2 {1}	-0,2489	0,5023	0,2076
3 {2}	-0,4831	0,2331	-0,6152
4 {2}	**-0,7163**	0,3325	-0,4098
5 {2}	**-0,7388**	0,2739	-0,1069
6 {2}	-0,5610	0,3352	-0,2193
7 {2}	0,3655	-0,0189	-0,1125
8 {2}	0,4680	-0,0965	0,1451

Variable	Factor Loadings (Unrotated) (Factor) Extraction: Principal components (Marked loadings are >,700000)		
	Factor_1	*Factor_2*	*Factor_3*
9 {1}	**-0,7396**	0,0771	0,4085
10 {1}	-0,4659	-0,0795	0,4612
11 {1}	**-0,7300**	0,3400	0,2463
12 {1}	-0,6324	0,0975	0,3207
13 {3}	**-0,7855**	-0,3697	-0,0776
14 {3}	**-0,7859**	-0,3814	-0,1056
15 {3}	**-0,7029**	-0,5632	-0,0270
16 {3}	**-0,7287**	-0,5560	-0,0213
Expl.Var	6,0577	1,9183	1,2716
Prp.Totl	0,3786	0,1199	0,0795

The eigenvalues for the model are presented in the Table V.

TABLE V. EIGENVALUES FOR THREE-FACTOR MODEL

Value	Eigenvalues (Factor) Extraction: Principal components			
	Eigenvalue	*% Total variance*	*Cumulative Eigenvalue*	*Cumulative %*
1	6,057701	37,86063	6,05770	37,8606
2	1,918289	11,98930	7,97599	49,8499
3	1,271599	7,94749	9,24759	57,7974

The three-factor model is significantly simpler than the first one (16 factors), but it can describe just ~58% of total variance. And even in this case the most meaningful correlation coefficients are observed for the 1st factor and for the different variable groups.

Let us provide the rotation of factors to get more acceptable meanings of factor loadings (it is equivalent to axes rotation). In this paper the Varimax raw method [12] is used. This method provides the maximization of the squared variance of original factor loadings for each factor variables (it is equivalent to variance maximization at the squared original factor loads matrix). The results are presented at the Table VI.

The rotation itself didn't improve the model quality, but it allowed to interpret it more clearly. Now, the 1st factor describes the educational result of the student (3rd group of characters), the 2nd factor is the qualification and innovation indicators of the school (1st group) and the 3rd factor is the social and economic conditions (2nd group). After the rotation, the contribution of each factor became more uniform that the last two rows of Table VI show.

TABLE VI. FACTOR LOADINGSS AFTER ROTATION

Variable	Factor Loadings (Varimax raw) (Factor) Extraction: Principal components (Marked loadings are >,700000)		
	Factor_1	*Factor_2*	*Factor_3*
1 {1}	-0,186816	0,505181	0,273903
2 {1}	-0,206110	0,514000	0,224968
3 {2}	0,181115	-0,061367	**0,793458**
4 {2}	0,259313	0,254851	**0,812035**
5 {2}	0,306132	0,451894	0,578273
6 {2}	0,144495	0,305678	0,600728
7 {2}	-0,230622	-0,285098	-0,110281
8 {2}	-0,253234	-0,196105	-0,383202
9 {1}	0,431374	**0,721672**	0,113635
10 {1}	0,357856	0,537752	-0,137395
11 {1}	0,238537	**0,725364**	0,355103
12 {1}	0,346814	0,611768	0,133154
13 {3}	**0,807891**	0,197652	0,260667
14 {3}	**0,817777**	0,172787	0,275040
15 {3}	**0,891417**	0,098338	0,087605
16 {3}	**0,903492**	0,119568	0,099939
Expl.Var	3,833031	2,808794	2,605764
Prp.Totl	0,239564	0,175550	0,162860

C. *Four-factor model*

Now we develop acceptable factor model, considering that this model will be more complicated. The 4th factor addition allowed describing just 65% of total variance. The rotation of the factors using Varimax method allowed to interpret clearly the factor loadings: the 1st factor is the educational results of student (the results of the USE), the 2nd factor is the school characteristics, the 3rd factor is student's family characteristics, the 4th factor is the teachers' characteristics.

Based on the research results we can conclude that for all types of schools the four-factor model describes just 65% of total variance. This is not sufficient for our study.

D. *Factor models for different types of schools*

First, we constructed the correlation matrix only for Tomsk urban schools. The most considerable (significantly deviated from 0) variables became the same 16 variables as for Tomsk region. The results of this investigation are presented in Table VII.

The results of the analysis have shown that the most simple three-factor model describes just about 67% of total variance for Tomsk schools. For Tomsk region, the results are 9% less. The four-factor model has the same results - 75% (65%).

TABLE VII. MODEL'S EIGENVALUES (TOMSK)

Value	Eigenvalues (Factor) Extraction: Principal components Include condition: V3=128 OR V3=129			
	Eigenvalue	*% Total variance*	*Cumulative Eigenvalue*	*Cumulative %*
1	7,640543	47,75340	7,64054	47,75340
2	1,556033	9,72521	9,19658	57,47860
3	1,476498	9,22811	10,67307	66,70672
4	1,286398	8,03998	11,95947	74,74670

The further investigation was made for the country, and ungraded schools. At first stage of the stochastic model development we calculated the correlation matrix to discover significant variables for each type of school. The results of the pair correlation analysis have shown that the most considerable variables became the same 16 variables as for Tomsk region (Table I). However, the calculated values of the correlation coefficients proved to be less than the same results for urban (Tomsk) schools. During further investigations the multifactor models were developed. The results of the analyses confirmed

our hypothesis about the low quality of the factor models for these types of schools. In fact, the best results for the four-factor model for country schools of Tomsk region described just about 40% of total variance.

III. CONCLUSION

During the research project the different factor models of Tomsk region schools functioning were developed and investigated using the software package STATISTICA [13]. At the first stage the most significant context factors (16 variables) were revealed. Using the method of principal components the exploratory data analysis was provided. Using the Cattell's and Kaiser-Guttman's criterions the optimal number of factors (5-6) was evaluated. The factor loadings analysis has shown that the most significant loadings s appeared at the first factor.

We investigated the three- and four-factor models as the simplest. The Varimax method for factor rotation allowed to interpret the characteristic groups according to the factor numbers clearly. It is observed that the simple models (3-4 factors) can not provide the required quality (describe less than 65% of total variance).

The model investigation for different types of schools has shown that for urban schools of Tomsk the quality of models is acceptable – 67% and 75% of total variance were described by the three- and four-factor models respectively.

The models developed for country and ungraded schools have very low quality due to the fact that the variables used in models have different effect on functioning quality of the different types of schools, and for country and ungraded schools the variables meanings have small effect on the quality characteristics of school functioning.

REFERENCES

[1]. V. A. Bolotov, V. A. Valdman, The conditions of the effective application the students' educational results estimation. [Pedagogics №6, 2012, pp. 39-45].

[2]. V. V. Kashpur, M. V. Rachilina, B. V. Ilyukhin, The background factors of the USE results. Tomsk: Deltaplan, 2008

[3]. V. A. Bolotov, V. A. Valdman, G. S. Kovaleva, M. A. Pinskaya, "The analysis of the Russian education quality assessment system constructing", in The educational management: theory and practice. Vol. 1, 2, 2011.

[4]. Yu. Ya. Katsman, A. V. Lepustin, B. V. Ilyukhin . The influence of contextual factors on the assessement of the effectiveness of work of schools in the Tomsk region]. [Vliyanie kontekstnyh faktorov na ocenku rezul'tatov effektivnosti raboty shkol tomskoy oblasti. Sovremennye problemy nauki i obrazovaniya,– N6., pp. 1-11, 2014. Available at: http://www.science-education.ru/120-16117.

[5]. B. V. Ilyukhin, A. V. Lepustin, Yu. Ya. Katsman. The mathematical modeling of the impact of context factors on the educational level of high school graduates in the Russian Federation. Proceedings of Tomsk State University of Control Systems and Radioelectronics [Matematicheskoe modelirovanie vliyaniya kontekstnyh faktorov na uroven' podgotovlennosti abiturientov uchrejdeniy vysshego professional'nogo obrazovaniya Rossiyskoy Federacii. Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki.], – Vol.2 (36), 2015, pp. 183-188. Available at: http://www.tusur.ru/filearchive/reports-magazine/2015-36-2/30.pdf.

[6]. V.V. Rybalko, The parametric diagnosing of the energetic objects based on the factor analyses using the STATISTICA environment. [Parametricheskoe diagnostirovanie energeticheskih ob"ektov na osnove faktornogo analiza v srede STATISTICA] Exponenta Pro., N2(6), 2004, pp. 78-83. Available at: http://www.statsoft.ru/solutions/ExamplesBase/branches/detail.php?ELEMENT_ID=643.

[7]. A. A. Halafyan STATISTICA 6. Statistical data analysis, 3rd ed. Tutorial. Moscow: Binom-press, 2007.

[8]. J. O. Kim, Ch. U. Muller, U. R. Klekk, Factor, discriminative and cluster analysis.Moscow: Finances and statistics, 1989.

[9]. I. T. Jolliffe, Principal component analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002

[10]. K. A. Yeomans and P. A. Golder, The Guttman-Kaiser criterion as a predictor of the number of common factors. Journal of the Royal Statistical Society. Series D (The Statistician) Vol. 31, No. 3 (Sep., 1982), pp. 221-229.

[11]. R. B. Cattell, Factor analysis. New York: Harper. 1952

[12]. H. F. Kaiser, The varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 1958, pp. 187–200.

[13]. J. Sá. Applied statistics using SPSS, STATISTICA, Matlab and R. Berlin: Springer, 2007.